European Society of Computational Methods in Sciences and Engineering (ESCMSE)

Journal of Numerical Analysis, Industrial and Applied Mathematics (JNAIAM)
vol. 19, no. 3-4, 2026, pp. 67-79
ISSN 1790–8140

Markov-Based Stochastic Analysis of Dengue Virus Sequences from the 2014 Japan Outbreak

Jesús E. García[1], V.A. González-López[2], D.S. Santos[3]

Department of Statistics,
Institute of Mathematics, Statistics and Scientific Computing,
University of Campinas,
Sergio Buarque de Holanda, 651, Campinas, S.P., CEP: 13083-859, Brazil

Abstract: In this paper, we investigate recently developed proximity measures within the framework of discrete-time Markov processes on discrete state spaces to assess, from multiple perspectives, the similarity and divergence among four Dengue Virus Type 1 genomic sequences available in FASTA format: LC011945, LC011948, LC011949, and LC016760, [1]. It has been speculated that these sequences may originate from multiple sources of serotype 1 of the virus, potentially leading to measurable discrepancies among them, [2]. Each sequence is modeled as a realization of a discrete-time Markov process of order 3 with alphabet {a, c, g, t}. Our analysis identifies LC011949 as the most dissimilar sequence. Building on this finding, we further identify the specific states most responsible for the observed discrepancies: taa (associated to a stop codon governing translation termination) and atc (related to isoleucine), in that order. We demonstrate that the proximity measures employed are effective in detecting subtle differences that are not immediately evident from the transition probabilities, which are also presented in our study. Based on this analysis, we propose a clustering structure: one group comprising sequences LC011945, LC011948, and LC016760, and a second group consisting solely of LC011949.

Keywords: Statistical consistency; Measures between samples; Bayesian information criterion, Efficient determination criterion

Mathematics Subject Classification: 60J10; 60J20

1 Introduction

Dengue virus is classified as an arbovirus and is maintained in nature through a transmission cycle involving mosquitoes of the Aedes genus, primarily Aedes aegypti. It belongs to the Flaviviridae family, which also includes the yellow fever virus - see [3]. There are four distinct serotypes of the dengue virus: DENV-1, DENV-2, DENV-3 and DENV-4, all of which are capable of causing both classic dengue fever and its more severe clinical manifestations, such as dengue hemorrhagic

[1]E-mail: jgarcia@iunicamp.br
[2]Corresponding author. E-mail: veronica@ime.unicamp.br
[3]d215862@dac.unicamp.br

fever and dengue shock syndrome. This potential for severe outcomes underscores the importance of continuous surveillance and monitoring of the virus's spread - [3]. Infection with one serotype confers lifelong immunity to that specific type but provides only temporary and partial cross-protection against the others. Subsequent infection with a different serotype increases the risk of severe disease, see [4]. Serotypes often compete for ecological dominance within a region, typically resulting in one or two serotypes becoming predominant for a certain period. Consequently, the local population may have little or no immunity to the non-dominant serotypes. If a previously uncommon or absent serotype is introduced into such a population, the risk of reinfection and severe clinical outcomes rises significantly. This dynamic of serotype circulation necessitates robust epidemiological monitoring and targeted public health interventions. When a particular serotype gains dominance in a region, it may spread to neighboring areas, altering the existing pattern of serotype prevalence. These shifting patterns of serotype circulation form the foundation and motivation for this research. Indeed, even within a single serotype, the dengue virus is known to exhibit enormous genetic variability [4, 5], which is also of epidemiological interest.

In this article, we apply tools from the theory of discrete stochastic processes to analyze the similarity and dissimilarity among four Dengue virus (DENV-1) sequences, provided in FASTA format, collected during the 2014 outbreak in Japan - see [1]. These genomic sequences are treated as realizations of underlying stochastic processes. By employing a range of similarity measures, we aim to quantify the degree of relatedness among the sequences and identify the one that appears most dissimilar, potentially indicating multiple entries of the serotype 1.

Beyond its immediate relevance to public health surveillance, identifying highly divergent Dengue virus sequences is of major clinical and industrial importance due to the virus's extensive genetic variability. Because partial cross-protection against newly emerging circulating strains can predispose individuals to severe secondary infections, pharmaceutical developers and diagnostic laboratories require precise tools to track dynamically shifting regions within the viral genome. In vaccine development, it is essential to determine whether reference-strain formulations remain effective against local outbreaks or if immune escape is occurring. Analyzing transition-probability dissimilarities provides a computationally lightweight alternative and complement to traditional multiple sequence alignment (MSA) techniques, see [6]. By capturing localized probabilistic shifts, this stochastic framework rapidly identifies non-homogeneous outbreak profiles, providing early signals to include emerging variants in updated vaccine formulations.

Our analysis builds upon the work of [2], which identified one sequence as notably divergent. However, while that study relied on a single dissimilarity measure, we take a broader approach by employing a family of dissimilarity measures, allowing for a potentially more nuanced and robust assessment of sequence relationships. In addition to reassessing which sequence stands out as an outlier, this study seeks to better understand the underlying causes of both similarity and divergence among the sequences. The use of diverse similarity measures offers deeper insight into the structural and probabilistic features that contribute to the observed patterns, improving our interpretation of the epidemiological dynamics involved.

To achieve our objective, we employ tools grounded in the Efficient Determination Criterion [7], applied to samples generated by stochastic processes. Our analysis focuses on identifying similarities and differences in the conditional stochastic behavior of the sequences. Formally, let $(W_t)_{t\in\mathbb{N}}$ denote a discrete-time Markov chain of finite order o over a finite and discrete alphabet, denoted by $\mathcal{G}$. The process has memory o, and its state space is given by the Cartesian product $\mathcal{G}^o = \mathcal{G}\times\cdots\times\mathcal{G}$ (with o factors). To denote realizations of the process, samples, states or specific configurations, we adopt the following notation: for $w_i \in \mathcal{G}$ and indices $k \le i \le m$, the concatenated sequence $w_k w_{k+1} \dots w_m$ is written as w_k^m. Since we are representing a stochastic process, an additional convention is required. In this article, sequences are read from left to right, with the rightmost element corresponding to the most recent value, w_m, and the leftmost element representing the most distant point in the past, w_k.

Returning to the problem at hand, we focus on sequences in FASTA format, where the underlying alphabet $\mathcal{G} = \{\text{a,c,g,t}\}$ corresponds to the four nucleotide bases. For a Markov chain of order $o = 3$, the state space becomes $\mathcal{G}^3 = \{\text{a, c, g, t}\} \times \{\text{a, c, g, t}\} \times \{\text{a, c, g, t}\}$, and each state $s \in \mathcal{G}^3$ can be identified with the concatenation of three elements from $\mathcal{G}$; for instance, agt denotes the state composed by the concatenation of a, g and t.
Once the value of o is known (or estimated), the stochastic process is fully identified by its sets of transition probabilities. For each $a \in \mathcal{G}$ and $s \in \mathcal{G}^o$, the transition probability from state s to symbol a is defined as

$$Q(a \mid s) = \mathbb{Q}(W_t = a \mid W_{t-o}^{t-1} = s). \tag{1}$$

The set of all such probabilities, $\{Q(a \mid s) : a \in \mathcal{G}, s \in \mathcal{G}^o\}$, uniquely determines the distribution of the process $(W_t)_{t\in\mathbb{N}}$. Thus, it is necessary to estimate $|\mathcal{G}|^o(|\mathcal{G}| - 1)$ parameters to fully determine $(W_t)_{t\in\mathbb{N}}$.
In order to estimate these transition probabilities, a finite sample (realization) of the process is required. Given a sample $w_1^n = w_1 w_2 \dots w_n$ of length n, generated by the stochastic process $(W_t)_{t\in\mathbb{N}}$, we define

$$C_n(s, a) = \left|\left\{t : o < t \le n, w_{t-o}^{t-1} = s, w_t = a \right\}\right|, \tag{2}$$

which is the number of occurrences of s immediately followed by a in the sample w_1^n. And, the total number of occurrences of the state s is

$$C_n(s) = \sum_{a\in\mathcal{G}} C_n(s, a). \tag{3}$$

Using these empirical counts, the transition probability defined in Equation (1) can be estimated via the maximum likelihood estimator:

$$\hat{Q}(a \mid s) = \frac{C_n(s, a)}{C_n(s)}, \quad \text{for } C_n(s) > 0. \tag{4}$$

The remainder of this article is organized as follows: Subsection 1.1 introduces the dataset; Subsection 1.2 presents the tools developed for analyzing Markov chains, along with the theoretical results that support their reliability. Section 2 reports the results, Section 3 provides the discussion, and Section 4 lists the references.
In the following subsection, we introduce the data under investigation.

1.1 Dengue Virus Type 1

Our study focuses on the complete genomes of four autochthonous DENV-1 (Dengue Virus Type 1), sequences isolated from patients during the 2014 outbreak in Japan (see [1]). The virus is transmitted to humans by infected mosquitoes of the Aedes genus. Dengue virus exists in four distinct serotypes and can infect individuals multiple times. This characteristic contributes to its high epidemiological efficiency. This virus's efficiency is one of the main reasons why there are numerous alerts when the dominant serotype changes in a given territory. This occurred in 2014 in Japan with serotype 1, and again in 2023 in Brazil with serotype 3.
The complete sequences were retrieved from the National Center for Biotechnology Information (NCBI) database at `http://www.ncbi.nlm.nih.gov/`. They were first sequenced and analyzed in [1]. We describe the sequences in Table 1.
In [1], it is speculated that the epicenter of the 2014 contamination was Yoyogi Park in Tokyo (patient 14-100J), but the question of multiple entries of the virus from neighboring countries also arises, and it is for this reason that we consider the four sequences described in Table 1.

Table 1: Complete sequences of Dengue Virus Type 1. From left to right: (1) the number of access to the NCBI base, (2) the patient from which the sequence is coming from, (3) the possible locale of contamination of the patient

Accesion number	Patient ID	Infected area
LC011945	14-100J	Yoyogi Park, Tokyo
LC011948	14-153J	Chiba
LC011949	14-181J	Shizuoka?
LC016760	14-188J	Hyogo? Malaysia?

Patient 14-153J had not visited Yoyogi Park for at least two weeks prior to the onset of dengue fever and was likely infected in Chiba Prefecture. Patient 14-181J, a resident of Shizuoka Prefecture, had not visited Yoyogi Park or any other known affected areas; however, he/she had traveled to other parts of Tokyo before symptom onset. Patient 14-188J lives in Nishinomiya City, Hyogo Prefecture, over 500 km from Tokyo, and had not traveled to the Tokyo area prior to developing symptoms. This patient had spent seven days in Malaysia and began exhibiting symptoms 12 days after returning.

The sequences we consider are in FASTA format, which facilitates our approach. We use the alphabet $\mathcal{G} = \{\text{a,c,g,t}\}$ with cardinal $|\mathcal{G}| = 4$. All the sequences have around a size of 10,700 elements. Heuristics provide a practical approach for eliciting the Markov chain order o. We consider a rule of thumb that relates the total number of transition probabilities - Equation (1), to be estimated, $|\mathcal{G}|^o(|\mathcal{G}| - 1)$, to the sample size n: $|\mathcal{G}|^o(|\mathcal{G}| - 1)\log(n) = n$. Under this criterion, approximately $\log(n)$ sample observations are allocated to estimate each transition probability. Solving for o, we obtain:

$$o = \frac{\log(n)}{\log(|\mathcal{G}|)} - \frac{\log(|\mathcal{G}| - 1)}{\log(|\mathcal{G}|)} - \frac{\log(\log(n))}{\log(|\mathcal{G}|)} \simeq \log_{|\mathcal{G}|}(n) - 1 - \frac{\log(\log(n))}{\log(|\mathcal{G}|)} < \log_{|\mathcal{G}|}(n) - 1,$$

or $o < \lfloor \log_{|\mathcal{G}|}(n) \rfloor - 1$ which in our case is $o < \lfloor \log_4(10,700) \rfloor - 1 = 5$. As commonly done in previous studies (e.g., [2]), we adopt $o = 3$, ensuring the interpretation in codons (compositions of 3 bases of $\mathcal{G}$) and ensuring that the sample size remains sufficient to yield reliable estimates of the transition probabilities (see Equations (1) and (4)).

In [2], the objective was to classify the four sequences in terms of their representativeness of the 2014 outbreak in Japan. That study also suggested that sequence LC011949 was the most divergent from the group, potentially indicating a distinct source of infection from the one responsible for the main outbreak. However, [2] did not investigate the underlying cause of this divergence. Motivated by this gap, we revisit the four sequences from a broader perspective, comparing different similarity criteria and identifying the specific state (or states) responsible for the observed discrepancies. The next subsection introduces the stochastic tool we use to compare the sequences.

1.2 The Family of Measures

First, let's state the context in which the tools we mention in this article operate. Consider two independent Markov chains $(W_{i,t})_{t\in\mathbb{N}}$ and $(W_{j,t})_{t\in\mathbb{N}}$, $i \neq j$, of finite order o, arranged on the finite alphabet $\mathcal{G}$ with state space $\mathcal{G}^o$. Given $s \in \mathcal{G}^o$ denote by $\{Q_i(a|s)\}_{a\in\mathcal{G}}$ and $\{Q_j(a|s)\}_{a\in\mathcal{G}}$ the sets of transition probabilities of $(W_{i,t})_{t\in\mathbb{N}}$ and $(W_{j,t})_{t\in\mathbb{N}}$, respectively - Equation (1). Consider also, two independent samples $w_{i,1}^{n_i}$, $w_{j,1}^{n_j}$ of $(W_{i,t})_{t\in\mathbb{N}}$ and $(W_{j,t})_{t\in\mathbb{N}}$, respectively. Define $C_{n_i+n_j}(s) = C_{n_i}(s) + C_{n_j}(s)$ and $C_{n_i+n_j}(s,a) = C_{n_i}(s,a) + C_{n_j}(s,a)$, $a \in \mathcal{G}$, where C_{n_i} and C_{n_j} are given as the samples w^{n_i} and w^{n_j}, respectively - see Equations (2) and (3).

In what follows, we introduce a formal framework for assessing the proximity or discrepancy between the set of probabilities $\{Q_i(a|s)\}_{a\in\mathcal{G}}$ and $\{Q_j(a|s)\}_{a\in\mathcal{G}}$ - Equation (1), using the empirical laws $\{\hat{Q}_i(a \mid s)\}_{a\in\mathcal{G}}$ and $\{\hat{Q}_j(a \mid s)\}_{a\in\mathcal{G}}$, see Equation (4).

Definition 1.1. *Given two independent Markov chains $(W_{i,t})_{t\in\mathbb{N}}$ and $(W_{j,t})_{t\in\mathbb{N}}$, of finite order o, arranged on the finite alphabet $\mathcal{G}$, and given two independent samples $w_{i,1}^{n_i}$, $w_{j,1}^{n_j}$ of $(W_{i,t})_{t\in\mathbb{N}}$ and $(W_{j,t})_{t\in\mathbb{N}}$, respectively, then for each state $s \in \mathcal{G}^o$, set*

$$d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = \frac{2}{(|\mathcal{G}|-1)\xi_{n_i+n_j}} \sum_{a\in\mathcal{G}} \Bigg\{ \sum_{k\in\{i,j\}} C_{n_k}(s,a) \ln\left(\frac{C_{n_k}(s,a)}{C_{n_k}(s)}\right) - C_{n_i+n_j}(s,a) \ln\left(\frac{C_{n_i+n_j}(s,a)}{C_{n_i+n_j}(s)}\right) \Bigg\},$$

where $\{\xi_{n_i+n_j}\}$ is a sequence of positive values indexed by $n_i + n_j$.

Each member of this family:

$$\{d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j})\}_{\xi_{n_i,n_j}},$$

holds several properties under adequate restrictions on $\xi_{n_i+n_j}$, see [8] and [9].

Definition 1.1 constitutes a measure on the space of samples and it is statistically consistent, meaning that, given the state $s \in \mathcal{G}^o$,

$$Q_i(a|s) = Q_j(a|s) \forall a \in \mathcal{G} \Leftrightarrow d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0; \tag{5}$$

$$\exists a \in \mathcal{G} : Q_i(a|s) \neq Q_j(a|s) \Leftrightarrow d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty. \tag{6}$$

To achieve such consistency, it is known that the sequence $\xi_{n_i+n_j}$ must satisfy certain conditions, such as

$$\frac{\xi_{n_i+n_j}}{\ln(\ln(n_i+n_j))} \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty, \quad \text{and} \quad \frac{\xi_{n_i+n_j}}{n_i+n_j} \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0, \tag{7}$$

or be of the form $\ln(\ln(n_i + n_j))$.
Since a global comparison between the sequences is sought, in order to expand the findings of [2], we introduce the following notion,

Definition 1.2. *Under the assumptions of Definition 1.1, set*

$$dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = \max_{s\in\mathcal{G}^o}\{d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j})\}.$$

Definition 1.2 preserves the consistency, as shown in the next result.

Theorem 1.1. *Under the assumptions of Definition 1.1, the notion introduced by Definition 1.2 verifies:*

(i) $Q_i(a|s) = Q_j(a|s) \forall a \in \mathcal{G}, \forall s \in \mathcal{G}^o \Leftrightarrow dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0.$

(ii) $\exists a \in \mathcal{G}, \exists s \in \mathcal{G}^o : Q_i(a|s) \neq Q_j(a|s) \Leftrightarrow dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty.$

Proof. To prove (i) and (ii) we will use Equations (5) and (6).
((i)⇒) note that under the assumption: $Q_i(a|s) = Q_j(a|s) \forall a \in \mathcal{G}, \forall s \in \mathcal{G}^o$ and applying Equation (5), $d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0$, $\forall s \in \mathcal{G}^o$, consequently $dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0.$

((i)⇐) under the assumption: $dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0$, and since $dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = d_{s^*}(w_{i,1}^{n_i}, w_{j,1}^{n_j})$, for at least one element $s^* \in \mathcal{G}^o$, we have that

$$0 \leq d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \leq d_{s^*}(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0, \forall s \in \mathcal{G}^o,$$

as a consequence, $d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} 0, \forall s \in \mathcal{G}^o \Rightarrow$ (applying Equation (5)) $Q_i(a|s) = Q_j(a|s) \forall a \in \mathcal{G}, \forall s \in \mathcal{G}^o$.
((ii)⇒) under the assumption: $\exists a \in \mathcal{G}, \exists s \in \mathcal{G}^o : Q_i(a|s) \neq Q_j(a|s)$, applying Equation (6), $d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty$, then it follows $dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty$.
((ii)⇐) from the assumption: $dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty$, there is at least one element $s^* \in \mathcal{G}^o$ such that $d_{s^*}(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j}) \underset{\min(n_i,n_j)\to\infty}{\longrightarrow} \infty$ and applying Equation (6), we have that $\exists a \in \mathcal{G} : Q_i(a|s^*) \neq Q_j(a|s^*)$, and the statement (ii) follows.

□

We further introduce the following concept, which aims to identify the state corresponding to the second maximum, thereby complementing the analysis of similarity and discrepancy among the sequences.

Definition 1.3. *Under the assumptions of Definition 1.1, set*

$$dmax(2)(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = \max_{s\in\mathcal{V}}\{d_s(w_{i,1}^{n_i}, w_{j,1}^{n_j})\},$$

with $\mathcal{V} = \left\{s : s \in \mathcal{G}^o \setminus \{s^* : d_{s^*}(w_{i,1}^{n_i}, w_{j,1}^{n_j}) = dmax(w_{i,1}^{n_i}, w_{j,1}^{n_j})\}\right\}$.

The following section shows the results of the comparisons.

2 Results

The objective is to detect differences between the four sequences, say $w_{1,1}^{n_1}$ (LC011945), $w_{2,1}^{n_2}$ (LC011948), $w_{3,1}^{n_3}$ (LC011949) and $w_{4,1}^{n_4}$ (LC016760). Then, we will treat the sequences listed in Table 1, expressed in FASTA format, as samples. The comparisons we propose are $w_{i,1}^{n_i}$ versus $w_{j,1}^{n_j}, i \neq j, i, j = 1, 2, 3, 4$ with $n_1 = n_2 = n_3 = n_4 = 10,693$.
In formulating the concepts presented in Definitions 1.1 and 1.2, it is necessary to fix the value of $\xi_{n_i+n_j}$, we do so according to the list provided in Table 2.

Table 2: Penalizations following conditions of Section 1.2. To the right of each term, $\xi_{n_i+n_j}$, the main reference is listed

$\xi_{n_i+n_j}$	Main reference
$\ln(n_i + n_j)$	[8]
$\{\ln(\ln(n_i + n_j))\}^2$	[10]
$\ln(\ln(n_i + n_j))$	[10]

Regarding the choice of penalty terms, we emphasize that the following will be used: (1): $\ln(\cdot)$, which satisfies the conditions in Equation (7) and is related to the Bayesian Information Criterion [8]; (2): $\ln(\ln(\cdot))$, as proposed in [10], related to the Efficient Determination Criterion. Although this second penalty violates the right-side condition in Equation (7), it still preserves all desirable

properties, as cited in Section 1.2. As discussed in [10], a wide range of penalty sequences is theoretically possible, however, in this study, we focus on penalties at the lower end of the spectrum, ranging from the minimal $\ln(\ln(\cdot))$ to the standard Bayesian Information Criterion penalty $\ln(\cdot)$ - [11]. The goal is to examine how smaller penalties influence the decision-making process, and because of that we also include in this study an intermediate penalty (3): $\ln(\ln(\cdot))^2$.

2.1 Similarity between the Sequences

In this section, we first present the *dmax* values for the sequence pairs $w_{1,1}^{n_1}$ (LC011945) - $w_{2,1}^{n_2}$ (LC011948), $w_{1,1}^{n_1}$ (LC011945) - $w_{3,1}^{n_3}$ (LC011949), $w_{1,1}^{n_1}$ (LC011945) - $w_{4,1}^{n_4}$ (LC016760), $w_{2,1}^{n_2}$ (LC011948) - $w_{3,1}^{n_3}$ (LC011949), $w_{2,1}^{n_2}$ (LC011948) - $w_{4,1}^{n_4}$ (LC016760) and $w_{3,1}^{n_3}$ (LC011949) - $w_{4,1}^{n_4}$ (LC016760), see Table 3. We also include dendrograms that illustrate the hierarchical organization of *dmax* values among the group of four sequences, see Figure 1. To construct each dendrogram, the *dmax* values are evaluated and ordered from smallest to largest. The height of the first horizontal branch (from bottom to top) joining the two closest sequences corresponds exactly to the numerical value of *dmax* at which this initial fusion occurs. The distance between the newly formed cluster and the remaining unmerged sequences is then updated using a linkage criterion. From these updated distances, the smallest value is identified to define the height of the second horizontal branch, forming a cluster that incorporates three of the four sequences. This iterative procedure continues until all sequences are fully integrated. Regarding the linkage criterion, we employed Complete Linkage, which considers the maximum *dmax* between elements of both clusters. Other linkage methods, such as Average and Single linkage, were also tested, yielding virtually identical hierar-chical structures. Furthermore, for each reported *dmax* value, we explicitly identify the specific state responsible for that maximum, as defined in Definition 1.2, see Table 3. As intended, these results are provided for the three penalty functions detailed in Table 2.

Table 3: *dmax* values - Definition 1.2, between the sequences $w_{1,1}^{n_1}$ (LC011945), $w_{2,1}^{n_2}$ (LC011948), $w_{3,1}^{n_3}$ (LC011949) and $w_{4,1}^{n_4}$ (LC016760), by pairs. In parentheses the reported state is the one that produces the highest d_s - Definition 1.1. From top to bottom with penalization (1) $\ln(n_i + n_j)$, (2) $\ln(\ln(n_i + n_j))^2$, (3) $\ln(\ln(n_i + n_j))$

$\ln(n_i + n_j)$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00153 (tgt)	0.05400 (taa)	0.00164 (ggt)
LC011948	-	0	0.05400 (taa)	0.00154 (cat)
LC011949	-	-	0	0.05400 (taa)
LC016760	-	-	-	0

$\ln(\ln(n_i + n_j))^2$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00289 (tgt)	0.10179 (taa)	0.00309 (ggt)
LC011948	-	0	0.10179 (taa)	0.00291 (cat)
LC011949	-	-	0	0.10179 (taa)
LC016760	-	-	-	0

$\ln(\ln(n_i + n_j))$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00664 (tgt)	0.23408 (taa)	0.00710 (ggt)
LC011948	-	0	0.23408 (taa)	0.00668 (cat)
LC011949	-	-	0	0.23408 (taa)
LC016760	-	-	-	0

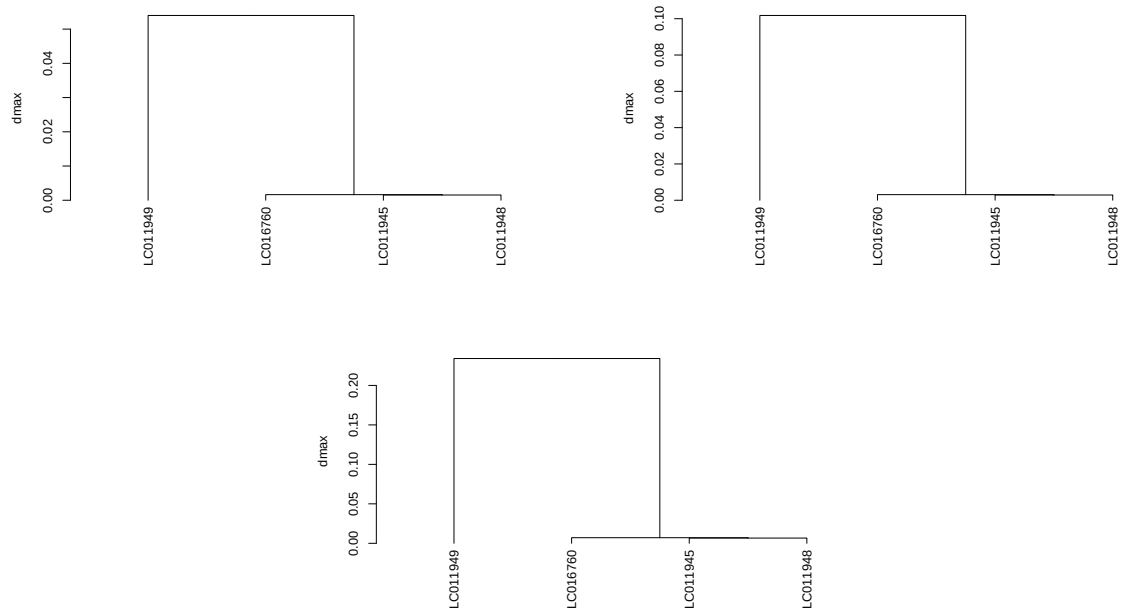

Figure 1: Dendrograms of the *dmax* values, reported in Table 3, for the sequences: LC011945, LC011948, LC011949 and LC016760. The vertical axis represents the numerical dissimilarity value (*dmax*) at which the clusters merge. From top to bottom, from left to right, with penalization (1) $\ln(n_i + n_j)$, (2) $\ln(\ln(n_i + n_j))^2$ and (3) $\ln(\ln(n_i + n_j))$, respectively

For each state responsible for the maximum value of d_s (according Table 3) we report the corresponding transition probabilities of the related sequences. Specifically, for state tgt, we present the transition probabilities of sequences LC011945 and LC011948 - Table 4; for state ggt, we report those of sequences LC011945 and LC016760 - Table 5; for state cat we report the transition probabilities of LC011948 and LC016760 - Table 6. Finally, Table 7 reports the transition probabilities across all sequences for state taa. This state yields the largest divergence, specifically separating sequence LC011949 from the other three.

Table 4: Estimated transition probabilities - Equation (4) - from the state tgt to the elements of $\mathcal{G}$

Sequence $w_{i,1}^{n_i}$	$\hat{Q}_i$(a | tgt)	$\hat{Q}_i$(c | tgt)	$\hat{Q}_i$(g | tgt)	$\hat{Q}_i$(t | tgt)
$(i=1), w_{1,1}^{n_1}$=LC011945	0.12658	0.17722	0.39241	0.30380
$(i=2), w_{2,1}^{n_2}$=LC011948	0.12579	0.17610	0.38365	0.31447

Table 5: Estimated transition probabilities - Equation (4) - from the state ggt to the elements of $\mathcal{G}$

Sequence $w_{i,1}^{n_i}$	$\hat{Q}_i$(a | ggt)	$\hat{Q}_i$(c | ggt)	$\hat{Q}_i$(g | ggt)	$\hat{Q}_i$(t | ggt)
$(i=1), w_{1,1}^{n_1}$=LC011945	0.18852	0.11475	0.40984	0.28689
$(i=4), w_{4,1}^{n_4}$=LC016760	0.18852	0.12295	0.40984	0.27869

Table 6: Estimated transition probabilities - Equation (4) - from the state cat to the elements of $\mathcal{G}$

Sequence $w_{i,1}^{n_i}$	$\hat{Q}_i$(a | cat)	$\hat{Q}_i$(c | cat)	$\hat{Q}_i$(g | cat)	$\hat{Q}_i$(t | cat)
$(i=2), w_{2,1}^{n_2}$=LC011948	0.21106	0.17085	0.43216	0.18593
$(i=4), w_{4,1}^{n_4}$=LC016760	0.20603	0.16583	0.43719	0.19095

Table 7: Estimated transition probabilities - Equation (4) - from the state taa to the elements of $\mathcal{G}$

Sequence $w_{i,1}^{n_i}$	$\hat{Q}_i$(a | taa)	$\hat{Q}_i$(c | taa)	$\hat{Q}_i$(g | taa)	$\hat{Q}_i$(t | taa)
$(i=1), w_{1,1}^{n_1}$=LC011945 $(i=2), w_{2,1}^{n_2}$=LC011948 $(i=4), w_{4,1}^{n_4}$=LC016760	0.35714	0.20408	0.21429	0.22449
$(i=3), w_{3,1}^{n_3}$=LC011949	0.40196	0.20588	0.14706	0.24510

In Table 8, we report the d_s values and associated states responsible for the second-highest maxima, excluding the primary maxima previously presented, see Definition 1.3. These values are provided under the three penalty scenarios described in Table 2.

Table 8: $dmax(2)$ values - Definition 1.3, between the sequences $w_{1,1}^{n_1}$ (LC011945), $w_{2,1}^{n_2}$ (LC011948), $w_{3,1}^{n_3}$ (LC011949) and $w_{4,1}^{n_4}$ (LC016760), by pairs. In parentheses the reported state is the one that produces the $dmax(2)$. From top to bottom with penalization (1) $\ln(n_i+n_j)$, (2) $\ln(\ln(n_i+n_j))^2$, (3) $\ln(\ln(n_i+n_j))$

$\ln(n_i+n_j)$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00058 (tga)	0.04383 (atc)	0.00126 (cat)
LC011948	-	0	0.04383 (atc)	0.00123 (tat)
LC011949	-	-	0	0.04515 (atc)
LC016760	-	-	-	0

$\ln(\ln(n_i+n_j))^2$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00109 (tga)	0.09091 (atc)	0.00238 (cat)
LC011948	-	0	0.08264 (atc)	0.00233 (tat)
LC011949	-	-	0	0.08513 (atc)
LC016760	-	-	-	0

$\ln(\ln(n_i+n_j))$	LC011945	LC011948	LC011949	LC016760
LC011945	0	0.00251 (tga)	0.19003 (atc)	0.00548 (cat)
LC011948	-	0	0.19003 (atc)	0.00535 (tat)
LC011949	-	-	0	0.19576 (atc)
LC016760	-	-	-	0

Figure 2 displays the dendrograms computed for the four sequences using the d_s, s =atc values (see Definition 1.1). The panels are arranged from left to right, from top to bottom, corresponding to the penalty scenarios described in Table 2. Note that state s =atc yields the maximum value $dmax(2)$ (see Definition 1.3) when comparing sequence LC011949 with all others. We highlight this state to show why sequence LC011949 exhibits the greatest discrepancy. In Table 9 we show, for each of the four sequences, the transition probabilities of the state atc for any element of the

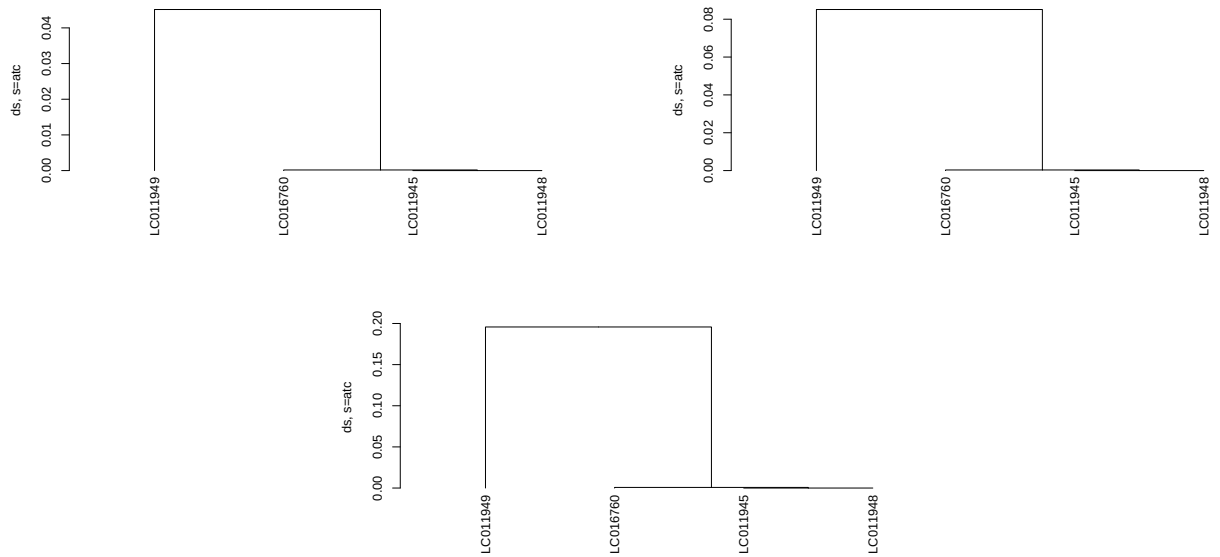

Figure 2: Dendrograms of the d_s, s =atc - Definition 1.1, reported in Table 8, for the sequences: LC011945, LC011948, LC011949 and LC016760. The vertical axis represents the numerical dissimilarity value (d_s) at which the clusters merge. From top to bottom, from left to right, with penalization (1) $\ln(n_i + n_j)$, (2) $\ln(\ln(n_i + n_j))^2$ and (3) $\ln(\ln(n_i + n_j))$, respectively

alphabet.

Table 9: Estimated transition probabilities - Equation (4) - from the state atc to the elements of $\mathcal{G}$

Sequence	$\hat{Q}_i$(a | atc)	$\hat{Q}_i$(c | atc)	$\hat{Q}_i$(g | atc)	$\hat{Q}_i$(t | atc)
$(i=1), w_{1,1}^{n_1}$=LC011945 $(i=2), w_{2,1}^{n_2}$=LC011948	0.40559	0.21678	0.09091	0.28671
$(i=4), w_{4,1}^{n_4}$=LC016760	0.40141	0.21831	0.09155	0.28873
$(i=3), w_{3,1}^{n_3}$=LC011949	0.41958	0.25175	0.09790	0.23077

Table 10 presents the states that produce the maximum d_s values ($dmax$ - Definition 1.2), along with those corresponding to the second-highest d_s values ($dmax(2)$ - Definition 1.3).

Table 10: States that produce the maximum d_s values - see Definition 1.1 - in the pairwise comparison of sequences: LC011945, LC011948, LC011949 and LC016760. Top: s_1, where $dmax = d_{s_1}$ - Definition 1.2; bottom: s_2, where $dmax(2) = d_{s_2}$ - Definition 1.3

s_1	LC011948	LC011949	LC016760
LC011945	tgt	taa	ggt
LC011948	-	taa	cat
LC011949	-	-	taa
s_2	LC011948	LC011949	LC016760
LC011945	tga	atc	cat
LC011948	-	atc	tat
LC011949	-	-	atc

3 Discussion

In [2], a comparative analysis was performed between the four sequences collected in Japan (LC011945, LC011948, LC011949 and LC016760) and all sequences available in the database `http://www.ncbi.nlm.nih.gov/`, that met the following criteria: complete Dengue Virus Type 1 genomes, collected in 2014, originating from Asian countries, specifically, China, Singapore, India, Malaysia, Sri Lanka and Thailand (see Table 5 and Figure 3 in [2]). The results indicate that sequence LC011949 displays a marked divergence from the other three sequences (LC011945, LC011948 and LC016760). In particular, these three sequences form a compact cluster in the dendrogram shown in Figure 3, whereas LC011949 is positioned at a substantially greater distance, underscoring its distinct evolutionary or structural profile. Nevertheless, it remains difficult to assert the existence of a clearly defined and easily recognizable pattern of geographic origin. Motivated by [2], we analyze the stochastic structures underlying the dissimilarities among the four Japanese sequences, and present our conclusions below.

Building on the results presented in Subsection 1.2 and in particular relying on the statistical consistency of Definition 1.1 and *dmax* (Definition 1.2 and Theorem 1.1), we present our findings in Section 2 under the assumption that a sample size of 10,700 is sufficient to ensure such consistency. Table 3 demonstrates that all three definitions of *dmax*, corresponding to different penalty functions (Definition 1.2 and Table 2), lead to consistent conclusions. In each case, the highest *dmax* value is observed when comparing sequence LC011949 with the remaining sequences (see Figure 1). Moreover, all three versions identify the same state: taa as responsible for this maximum value. When comparing the *dmax* values across different penalty levels, from the highest penalty (on top row of Table 3) to the lowest (bottom row of Table 3), we observe that the magnitude of *dmax* increases proportionally as the penalty decreases. Specifically, the values at the bottom of the table are approximately four times greater than those at the top, for example, $dmax$(LC011945,LC011949) = 0.05400 with the highest penalty, it increases to $dmax$(LC011945,LC011949) = 0.23408 with the lowest penalty, and $0.23408 \approx 4 \times 0.05400$. This trend is also visually evident in Figure 1, where the vertical axis clearly shows the increasing magnitudes from left to right, and from top to bottom. It should be noted that a reduction in the penalty term may permit the application of the proposed procedure to datasets with smaller sample sizes. This hypothesis is currently being assessed through simulations and therefore falls outside the scope of the present study. Nonetheless, preliminary evidence suggests that, if validated, the adoption of a ln(ln())-based penalty would provide an adequate choice.

As reported in Table 3, the states that lead to the maximum values of d_s are four: tgt, ggt, cat, and taa, consistently across the three penalties shown in Table 2. The transition probabilities starting

from each of the three states: tgt, ggt, and cat, for each pair of sequences involved, LC011945 and LC011948 in the case of tgt, LC011945 and LC016760 in the case of ggt, and LC011948 and LC016760 in the case of cat, are quite close, as shown in Tables 4, 5 and 6, respectively. These small variations result in $dmax$ values of up to 0.002 (using $\ln(\cdot)$), 0.003 (using $\ln(\ln(\cdot))^2$), and 0.007 (for $\ln(\ln(\cdot))$).

The results reported in Section 2 show that among the four states identified as contributing to the maximum d_s values, taa stands out, since the transition probabilities associated with sequences LC011945, LC011948 and LC016760 are identical, and they differ from those corresponding to sequence LC011949. This fact is detailed in Table 7 and illustrated clearly in Figure 1, which highlights the relevance of the taa state, as it consistently accounts for the maximum d_s values in all comparisons involving LC011949 and the other sequences. In the universal genetic code, the triplet taa is associated to a stop codon governing translation termination, while also frequently occurring within the non-coding regions ($5'$ UTR and $3'$ UTR) that regulate viral replication and host-pathogen interactions [13].

A closer examination of the state associated with the second-highest d_s value, excluding the $dmax$ case - Definition 1.3, reveals that, once again, sequence LC011949 stands out as the most different from the other three sequences - see Table 8 and Figure 2. This pattern holds for a single state: atc, across all three penalty scenarios - see Table 2. The triplet atc is associated to one of three codons that signal the ribosome to insert an isoleucine residue into the emerging viral polyprotein. Table 9 and Figure 2 (conditional to state atc) are best interpreted in conjunction. As observed previously, sequence LC011949 is clearly distinguished from the other three. Moreover, sequence LC016760 exhibits transition probability values that differ slightly from those of the remaining two sequences, LC011945 and LC011948 - see Table 9.

By analyzing Table 10 alongside Tables 3 and 8, and Figures 1 and 2, a more precise assessment can be made of the extent to which sequence LC011949 differs from the other three. This analysis also highlights the states with the greatest discrepancies, namely, taa and atc.

The extension of the studies conducted in [2], as carried out in this paper, enables the identification of two distinct groups of sequences: one comprising: LC011945, LC011948 and LC016760, and the other consisting solely of sequence LC011949. Moreover, we highlight that using smaller penalties, such as the $\ln(\ln(\cdot))$ function, can yield conclusions consistent with those obtained in [2], while easing the sample size requirements typically imposed by larger penalties.

Beyond epidemiological surveillance, the stochastic identification of specific divergent state transition probabilities, such as taa and atc, provides valuable tools for bioinformatics and ratio-nal vaccine design, particularly within DNA/RNA-based vaccine platforms. By capturing subtle, alignment-free probabilistic shifts across viral genomes, this framework may enable the rapid profiling of localized codon usage patterns and structural motifs without the computational burden of multiple sequence alignment. Detecting such divergences may allow bioinformaticians to better optimize synthetic gene constructs, refine codon selection, and design targeted or broad-spectrum formulations that effectively account for emerging viral variants.

Acknowledgment

D. S. Santos gratefully acknowledge the financial support provided by CAPES with fellowships from the PhD Program in Statistics, University of Campinas. The author wishes to thank the anonymous referees for their careful reading of the manuscript and their fruitful comments and suggestions.

References

[1] S. Tajima, E. Nakayama, A. Kotaki, M.L. Moi, M. Ikeda, K. Yagasaki, Y. Saito, K. Shibasaki, M. Saijo and T. Takasaki, Whole-genome sequencing-based molecular epidemiologic analysis of autochthonous dengue virus type 1 strains circulating in Japan in 2014, *Japanese Journal of Infectious Diseases* **70** 45-49(2017).

[2] M.T.A. Cordeiro, J.E. García, V.A. González-López and S.L.M. Londoño, Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014, *4open* **2** 20(2019).

[3] E.A. Henchal and J.R. Putnak, The dengue viruses, *Clin. Microbiol. Rev.* **3** 376-396(1990).

[4] C.P. Simmons, J.J. Farrar, N.v.V. Chau and B. Wills, Dengue, *N. Engl. J. Med.* **366** 1423-1432(2012).

[5] E.C. Holmes and S.S. Twiddy, The origin, emergence and evolutionary genetics of dengue virus, *Infect. Genet. Evol.* **3** 19-28(2003).

[6] K. Muthumani, D.J. Laddy, B.N. Shukla, K.M. Lankaraman and D.B. Weiner, Novel synthetic consensus-based DNA vaccine encoding Dengue virus envelope protein domain III induces broad humoral and cellular immune responses, *Vaccine* **26** 5186-5193(2008).

[7] L.C. Zhao, C.C.Y. Dorea and C.R. Gonçalves, On determination of the order of a Markov chain, *Statistical Inference for Stochastic Processes* **4** 273-282(2001).

[8] J.E. García, R. Gholizadeh and V.A. González-López, A BIC-based consistent metric between Markovian processes, *Appl. Stochastic Models Bus. Ind.* **34** 868-878(2018).

[9] D.F.S. Pereira: *Efficient Determination Criterion for Estimation of Minimal Partition Markov Chains*. MSc Dissertation, University of Brasilia, Brazil, 2021 (in Portuguese).

[10] J.E. García, V.A. González-López and J.I. Gomez Sanchez, A metric based on the Efficient Determination Criterion, *Entropy* **26** 526(2024).

[11] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics* **6** 461-464(1978).

[12] J.E. García and V.A. González-López, Consistent estimation of Partition Markov Models, *Entropy* **19** 160(2017).

[13] T.J. Chambers, C.S. Hahn, R. Galler and C.M. Rice, Flavivirus genome organization, expression, and replication, *Annu. Rev. Microbiol.* **44** 649-688(1990).